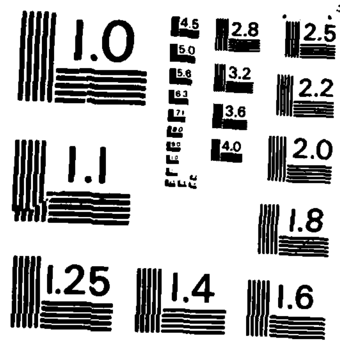
AD-A166 897 ALL SUBSETS REGRESSION IN COX MODEL(U) STANFORD UNIV CA DEPT OF STATISTICS A Y KUK 04 MAR 86 TR-370 N00014-76-C-0475 1/1

UNCLASSIFIED F/G 12/1 NL




MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

# ALL SUBSETS REGRESSION IN COX MODEL

BY

ANTHONY Y.C. KUK

TECHNICAL REPORT NO. 370
MARCH 4, 1986

PREPARED UNDER CONTRACT
N00014-76-C-0475 (NR-042-267)
FOR THE OFFICE OF NAVAL RESEARCH



DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

# ALL SUBSETS REGRESSION IN COX MODEL

BY

ANTHONY Y.C. KUK

TECHNICAL REPORT NO. 370

~~FEBRUARY 27~~, 1986
MARCH 4,

Prepared Under Contract
N00014-76-C-0475 (NR-042-267)
For the Office of Naval Research

Herbert Solomon, Project Director





DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

# 1. INTRODUCTION

Stepwise procedures are often used to select concomitant variables in regression with censored data (Krall, Uthoff & Harley, 1975; Peduzzi, Hardy & Holford, 1980; Lee, Harrel, Tolley & Rosati, 1983). An undesirable feature of stepwise procedures is that they lead to a single subset of variables and do not suggest alternative good subsets. Another concern is the possibility of premature termination. We shall see later that this is indeed the case when stepwise procedures are applied to the multiple myeloma data (Krall, Uthoff & Harley, 1975). In comparison, all subsets regression provides more information, is more reliable and is to be preferred provided that it is computationally feasible. The purpose of this paper is to show that within the framework of the proportional hazards model (Cox, 1972) all subsets regression can be performed with very little computational efforts.

The first selection criterion that we consider is based on cross-validation. We argue that the correct way to cross-validate in our setting is to change the status of one observation from uncensored to censored. An asymptotically equivalent criterion that requires less computation is the following: if $\alpha$ indexes the model, choose $\alpha$ to minimize

$$\Lambda_\alpha = W_\alpha + 2P_\alpha , \tag{1}$$

where $W_\alpha$ denotes the partial likelihood ratio statistic for testing the model $\alpha$ against the full model and $P_\alpha$ is the number of covariates included in model $\alpha$. The criterion that requires least computation is

$$\Lambda'_\alpha = W'_\alpha + 2P_\alpha \quad , \tag{2}$$

where $W'_\alpha$ denotes Wald statistic. We show that criterion $\Lambda'$ is formally equivalent to Mallow's $C_p$. As a result, we are able to compute and compare the values of $\Lambda'$ for all possible subsets making use of standard statistical packages. We apply this to the multiple myeloma data and obtain results remarkably different from those obtained by previous workers using stepwise procedures. New insights are gained and the superiority of all subsets regression over stepwise regression is clearly demonstrated.

## 2. CRITERIA FOR SELECTING VARIABLES

The proportional hazards model is specified by the hazard relationship

$$\lambda(t;x) = \lambda_0(t)\exp(x\beta)$$

where x is a row vector of P covariates, $\beta$ is a column vector of P regression constants and $\lambda_0(t)$ is an arbitrary and unspecified baseline hazard function. Let $(t_j,\delta_j)$, $j = 1,\ldots,n$ be an observed sample of failure times with $\delta_j = 0$ indicating a right-censored observation and $\delta_j = 1$ indicating a failure. The associated covariate vectors are $x_1, \ldots, x_n$. An estimate of $\beta$ is obtained by maximizing the partial likelihood

$$L(\beta) = \prod_{i=1}^{k} \frac{e^{x_{(i)}\beta}}{\sum_{j\in R_{(i)}} e^{x_j\beta}} , \tag{3}$$

where $t_{(1)} < \ldots < t_{(k)}$ are the uncensored failure times with corresponding covariates $x_{(1)}, \ldots, x_{(k)}$, censoring status $\delta_{(1)}, \ldots, \delta_{(k)}$ and $R_{(i)}$ is the set of individuals known to be alive just prior to $t_{(i)}$. By setting different components of $\beta$ to zero, we obtain $2^P$ submodels of the full model.

We propose the following criterion for model choice: if $\alpha$ indexes the model, choose $\alpha$ to maximize

$$\prod_{i=1}^{k} \frac{e^{x_{(i)}\hat{\beta}_{-i}(\alpha)}}{\sum_{j\in R_{(i)}} e^{x_j\hat{\beta}_{-i}(\alpha)}} , \tag{4}$$

where $\hat{\beta}_{-i}(\alpha)$ denotes the maximum partial likelihood estimate of $\beta$ computed under model $\alpha$ when $\delta_{(i)}$ is changed from one to zero. We call (4) the cross-validatory criterion for two reasons. Firstly, we note that changing $\delta_{(i)}$ from one to zero corresponds to removing the ith term from the partial likelihood (3). This is similar mathematically to ordinary cross-validation where the effect of deleting one observation is to remove one term from the likelihood function. Secondly, we note that the ith term of (3) is the conditional probability

$$P\{\text{death of (i) at time } t_{(i)} \mid \text{one death in } R_{(i)} \text{ at time } t_{(i)}\}. \quad (5)$$

It would be unrealistic to access the choice of $\alpha$ with

$$\prod_{i=1}^{k} \frac{e^{x_{(i)}\hat{\beta}(\alpha)}}{\sum_{j\varepsilon R_{(i)}} e^{x_j\hat{\beta}(\alpha)}} \; . \quad (6)$$

Since the information that (i) died at $t_{(i)}$ is used to obtain $\hat{\beta}(\alpha)$, the ith term of (6) is a biased estimate of (5). If we change $\delta_{(i)}$ from one to zero, we still retain the information that (i) is alive just prior to $t_{(i)}$ but the fact that (i) died at $t_{(i)}$ is no longer used. This is very similar in spirit to ordinary cross-validation.

While the cross-validatory criterion is of theoretical interest, its computation is prohibitive. For each of the $2^P$ models, we have to compute $\hat{\beta}_{-i}$, $i = 1,\ldots,k$, each of which requires iteration. By following the proof of Stone (1977), we can show the asymptotic equivalence of model choice by cross-validation and criterion $\Lambda$. In particular, if $\alpha_1 \subset \alpha_2$ and $\alpha_1$ is correct, then the asymptotic significance level of either criterion is $P(\chi^2_\nu > 2\nu)$ where $\nu = P_{\alpha_2} - P_{\alpha_1}$. Criterion $\Lambda$ as defined in (1) requires less computation since we only need to compute one $\hat{\beta}$ for each model. In spite of this, the task remains formidable. An alternative criterion is $\Lambda'$ which is based on Wald statistic. Peace & Flora (1978), Lee, Harrel, Tolley & Rosati (1983) compare the partial likelihood ratio statistic and Wald statistic and find them comparable in accessing the effects of concomitant variables in survival analysis. We also find good agreement between $\Lambda$ and $\Lambda'$. Criterion $\Lambda'$ as defined in (2) is computationally simplier. If we partition $\beta^T$ as $(\beta_1^T, \beta_2^T)$ and $\hat{\beta}^T = (\hat{\beta}_1^T, \hat{\beta}_2^T)$ is obtained under the full model, then $W'_\alpha = \hat{\beta}_2^T C_{22}^{-1} \hat{\beta}_2$ where without loss of generality we have assumed that model $\alpha$ corresponds to setting $\beta_2 = 0$ and

$$C = \begin{bmatrix} C_{11} & C_{12} \\ C_{22} & C_{22} \end{bmatrix} = I^{-1} \quad ,$$

the inverse of the information matrix, is the estimated covariance matrix of $\hat{\beta}$. Lawless & Singhal (1978) note that if we begin with

$$\begin{bmatrix} I & I^T\hat{\beta} \\ \hat{\beta}^T I & \hat{\beta}^T I \hat{\beta} \end{bmatrix} \quad , \qquad (7)$$

then by operating on the matrix with a sequence of sweep operations, we can obtain $W'_\alpha = \hat{\beta}_2 C_{22}^{-1} \hat{\beta}_2$ for all $2^P$ models. Instead of (7), we use the matrix

$$\begin{bmatrix} I & I^T\hat{\beta} \\ \hat{\beta}^T I & (N-P-1) + \hat{\beta}^T I \hat{\beta} \end{bmatrix}, \qquad (8)$$

where N is an arbitrary integer greater than P. This matrix can be obtained easily since $\hat{\beta}$ and $C = I^{-1}$ are the standard output of any program that does Cox regression. The program that we use is the BMDP program P2L. If we treat (8) as if it were the matrix of corrected sums of squares and cross products of the independent and dependent variables computed from a sample of size N, then criterion $\Lambda'$ is formally equivalent to Mallow's $C_p$. To see this, note that

$$C_p(\alpha) = \frac{RSS(\alpha)}{\hat{\sigma}^2} + 2(P_\alpha + 1) - N$$

where $\hat{\sigma}^2 = RSS(\text{full model})/(N - P - 1) = 1$ by our choice of (8). Since $RSS(\alpha) = RSS(\text{full model}) + \hat{\beta}_2^T C_{22}^{-1} \hat{\beta}_2$,

$$C_p(\alpha) = \Lambda'_\alpha - (P - 1)$$

and the two criteria are equivalent. The problem can now be handled by standard statistical packages. We use the BMDP program P9R which does all subsets linear regression.

## 3. AN EXAMPLE

Krall, Uthoff & Harley (1975) presents a data set consisting of the survival times of sixty five multiple myeloma patients with sixteen concomitant variables. Seventeen of the observations are censored. They assume an exponential regression model in which the mean survival time is a linear function of the concomitant variables. A step-up procedure based on the likelihood ratio criterion leads to the subset {1,2,16}. Lawless & Singhal (1978) adopt a proportional hazards exponential regression model

$$\lambda(x) = \lambda_0 \exp(x\beta) \ .$$

They report the best three subsets of each size, according to the value of the likelihood ratio statistic and Wald statistic. The best subset of size two is {1,2}. To compare subsets of different sizes, we use Akaike's criterion and end up with {1,2} as the best subset. Unfortunately, Lawless & Singhal do not consider all sixteen concomitant variables of the original data set. Instead, they consider only variables 1, 2, 3, 5, 6, 7, 9, 16. We shall see later that this is a very poor choice. Peduzzi, Hardy & Holford (1980) also assume a multiplicative exponential regression model and use a stepwise procedure to identify {1,2} as the best subset of the eight variables considered by Lawless & Singhal. Their criterion for inclusion of a variable is based on the score statistic for testing the current model against the candidate model. The criterion for removal of a variable is based on Wald statistic. The above analyses are all based on exponential regression models, an analysis based on Cox model appears as an example in BMDP manual where stepwise regression based on the partial likelihood ratio criterion leads again to the subset {1,2}. Our analysis provides a lot more information than any stepwise procedure. We report in Table 1 the best three subsets of each size according to $CP = \Lambda' - 15$. For comparison purpose, we also compute the values of $LR = \Lambda - 15$ for these models. The agreement between $\Lambda$ and $\Lambda'$ appears to be good especially for the better subsets. It can be observed that the value of CP for the best subset of size $P_\alpha$ decreases initially, reaches a bottom at $P_\alpha = 8$, and then begin to increase. The best subset is {1,2,4,6,7,8,12,13} which has a log partial likelihood of -140.20 compared with -138.14 of the full model. The relationship among the best five subsets is shown in Figure 1 and summary statistics for the best subset is given in Table 2. The subset {1,2} selected by the majority of previous workers is far from being the best. Its CP value of 11.04 and LR value of 10.52 are considerably larger than the corresponding values 5.00 and 5.11 of the best subset. The associated log partial likelihood of -148.90 is unimpressive. The partial likelihood ratio statistic for testing the subset {1,2} against the best subset is 17.4 which is significant at the 0.01 level. Neverthless, we observe that the addition of an extra variable to {1,2} does not add much. In fact, the improvement is gradual until we reach $P_\alpha = 7$. This explains the selection of {1,2} by stepwise procedures. Lastly, we consider {1,2,3,5,6,7,9,16}, the set of variables used by Lawless & Singhal. Its CP value of 17.84 and LR value of 17.66 are large. The associated log partial likelihood is -146.47 compared with -140.20 of the best subset which also has eight variables and -146.54 of the best subset of size 4.

# REFERENCES


COX, D.R. (1972). Regression models and life-tables (with discussion). *J. R. Statist. Soc.* B **34**, 187-202.

KRALL, J.M., UTHOFF, V.A. & HARLEY, J.B. (1975). A step-up procedure for selecting variables associated with survival. *Biometrics* **31**, 49-57.

LAWLESS, J.F. & SINGHAL, K. (1978). Efficient screening of nonnormal regression models. *Biometrics* **34**, 318-27.

LEE, K.L., HARRELL, F.E., TOLLEY, H.D. & ROSATI, R.A. (1983). A comparison of test statistics for assessing the effects of concomitant variables in survival analysis. *Biometrics* **39**, 341-50.

PEACE, K.E. & FLORA, R.E. (1978). Size and power assessments of tests of hypotheses on survival parameters. *J. Am. Statist. Assoc.* **73**, 129-32.

PEDUZZI, P.N., HARDY, R.J. & HOLFORD, T.R. (1980). A stepwise variable selection procedure for nonlinear regression models. *Biometrics* **36**, 511-6.

STONE, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. R. Statist. Soc.* B **39**, 44-7.

Table 1. Best three models of each size.

| $P_\alpha$ | Variables in model | Overall Position | CP | LR | Log partial likelihood |
|---|---|---|---|---|---|
| 1 | 1 | | 13.28 | 12.67 | -150.98 |
| | 2 | | 14.35 | 15.61 | -152.45 |
| | 13 | | 16.71 | 18.39 | -153.84 |
| 2 | *1,2 | | 11.04 | 10.52 | -148.90 |
| | 1,13 | | 12.22 | 11.52 | -149.40 |
| | 1,4 | | 13.92 | 13.98 | -150.63 |
| 3 | 1,2,16 | | 10.93 | 10.82 | -148.05 |
| | 1,2,13 | | 10.96 | 10.77 | -148.03 |
| | 1,2,14 | | 11.11 | 11.02 | -148.16 |
| 4 | 1,3,12,13 | | 9.26 | 9.80 | -146.54 |
| | 1,12,13,14 | | 10.15 | 9.22 | -146.25 |
| | 1,2,12,13 | | 10.23 | 10.63 | -146.96 |
| 5 | 1,2,12,13,14 | | 7.99 | 8.63 | -144.96 |
| | 1,3,7,12,13 | | 8.59 | 10.17 | -145.73 |
| | 1,3,4,12,13 | | 8.89 | 9.84 | -145.56 |
| 6 | 1,3,4,7,12,13 | | 7.96 | 9.28 | -144.28 |
| | 1,2,7,12,13,14 | | 8.00 | 9.12 | -144.20 |
| | 1,3,6,7,12,13 | | 8.10 | 8.95 | -144.12 |
| 7 | 1,3,4,6,7,12,13 | 5 | 6.45 | 7.59 | -142.44 |
| | 1,3,4,7,8,12,13 | | 7.11 | 7.67 | -142.48 |
| | 1,3,4,7,12,13,14 | | 8.04 | 9.19 | -143.24 |
| 8 | 1,3,4,6,7,8,12,13 | 1 | 5.00 | 5.11 | -140.20 |
| | 1,3,4,7,8,12,13,14 | | 7.23 | 7.65 | -141.47 |
| | 1,2,3,4,6,8,12,13 | | 7.28 | 8.18 | -141.73 |
| | †1,2,3,5,6,7,9,16 | | 17.84 | 17.66 | -146.47 |
| 9 | 1,2,3,4,6,7,8,12,13 | 2 | 5.89 | 5.85 | -139.57 |
| | 1,3,4,6,7,8,12,13,14 | 3 | 6.15 | 6.15 | -139.72 |
| | 1,3,4,5,6,7,8,12,13 | | 6.53 | 6.75 | -140.02 |
| 10 | 1,2,3,4,6,7,8,12,13,14 | 4 | 6.33 | 6.32 | -138.80 |
| | 1,3,4,6,7,8,12,13,14,15 | | 7.08 | 7.16 | -139.22 |
| | 1,2,3,4,6,7,8,12,13,15 | | 7.66 | 7.64 | -139.46 |
| 11 | 1,2,3,4,6,7,8,12,13,14,15 | | 7.23 | 7.23 | -138.26 |
| | 1,2,3,4,5,6,7,8,12,13,16 | | 9.17 | 9.32 | -139.30 |
| | 1,3,4,5,6,7,8,10,11,12,13 | | 10.15 | 10.33 | -139.81 |
| 12 | | | 9.13 | 9.12 | -138.21 |
| 13 | | | 11.07 | 11.07 | -138.18 |
| 14 | | | 13.02 | 13.02 | -138.15 |
| 15 | | | 15.01 | 15.01 | -138.15 |
| Full model | | | 17.00 | 17.00 | -138.14 |

* This is the subset selected by stepwise procedures.

† This is Lawless & Singhal's subset.

$$
\begin{tikzcd}
P_\alpha = 7 & & \{1, 3, 4, 6, 7, 12, 13\} \arrow[d, "+\{8\}"] & \\
P_\alpha = 8 & & \{1, 3, 4, 6, 7, 8, 12, 13\} \arrow[dl, "+\{2\}"'] \arrow[dr, "+\{14\}"] & \\
P_\alpha = 9 & \{1, 2, 3, 4, 6, 7, 8, 12, 13\} \arrow[dr, "+\{14\}"'] & & \{1, 3, 4, 6, 7, 8, 12, 13, 14\} \arrow[dl, "+\{2\}"] \\
P_\alpha = 10 & & \{1, 2, 3, 4, 6, 7, 8, 12, 13, 14\} &
\end{tikzcd}
$$

Fig. 1. Relationship among the best five subsets

Table 2. Summary statistics for the best subset.

Log partial likelihood = -140.20

| Variable | Coefficient | Standard error | Coeff./S.E. | Exp(Coeff.) |
|---|---|---|---|---|
| 1 | 1.9160 | .6112 | 3.1346 | 6.7938 |
| 3 | -1.5439 | .5025 | -3.0726 | 0.2135 |
| 4 | .9889 | .4367 | 2.2647 | 2.6883 |
| 6 | -.8171 | .3951 | -2.0681 | 0.4417 |
| 7 | 1.8418 | .7701 | 2.3915 | 6.3081 |
| 8 | .8047 | .4078 | 1.9734 | 2.2360 |
| 12 | .1070 | .0311 | 3.4384 | 1.1130 |
| 13 | 1.5074 | .4147 | 3.6345 | 4.5149 |

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

# REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

| | |
|---|---|
| 1. REPORT NUMBER 370 | 2. GOVT ACCESSION NO. AD-A166 897 |
| 3. RECIPIENT'S CATALOG NUMBER | |
| 4. TITLE (and Subtitle) All Subsets Regression In Cox Model | 5. TYPE OF REPORT & PERIOD COVERED TECHNICAL REPORT |
| | 6. PERFORMING ORG. REPORT NUMBER |
| 7. AUTHOR(s) Anthony Y.C. Kuk | 8. CONTRACT OR GRANT NUMBER(s) N00014-76-C-0475 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, CA 94305 | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-042-267 |
| 11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistics & Probability Program Code 411SP | 12. REPORT DATE ~~February 27~~ MARCH 4, 1986 |
| | 13. NUMBER OF PAGES 15 |
| 14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) | 15. SECURITY CLASS. (of this report) UNCLASSIFIED |
| | 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE |

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Cross-validation; Mallow's $C_p$; Proportional hazards model; Wald statistic.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

PLEASE SEE FOLLOWING PAGE.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

TECHNICAL REPORT NO. 370

20. ABSTRACT

This paper shows that within the framework of the proportional hazards model all subsets regression can be performed with very little computational efforts. A selection criterion based on Wald statistic is motivated by a cross-validation argument in which the status of one observation is changed from uncensored to censored. This criterion is seen to be formally equivalent to Mallow's $C_p$ and thus the problem is reduced to one readily handled by standard statistical packages. The procedure is applied to the multiple myeloma data to give results remarkably different from those obtained by previous workers using stepwise procedures. New insights are gained and the superiority of all subsets regression over stepwise regression is clearly demonstrated.

END

DTIC

5-86